Appendix

This supplementary provides additional details and results that could not fit in the main paper. Specifically, we include: (1) dataset and annotation generation (Gibson and HM3D), (2) top-down visualization of co-visibility graph, (3) evaluation metric details, (4) graph definitions used in DUSt3R and CroCo experiments, (5) ablation studies on Covis, and (6) Sim2Real downstream task details. All experiments and dataset generation were conducted on an HPC cluster with A100 and V100 GPUs, using a single GPU per run.

A. Dataset details

We detail the dataset construction and annotation processes for the Gibson and HM3D environments, which form the basis of our co-visibility analysis. For each dataset, we define *scenarios* as structured collections of images with associated camera poses, designed to ensure high coverage and annotation quality. Both datasets leverage simulated sensors (pinhole and stereo) to generate RGB and depth data, and we outline specific strategies used to handle scene diversity, pose recording, and structural alignment.

Gibson: This dataset covers 129 floors across 85 scenes. To ensure diversity, we randomize seed numbers during data generation, allowing for efficient image selection that maximizes spatial coverage with minimal redundancy. The image selection process is further detailed in Sec. B.1. We record precise camera poses using Habitat-sim and organize images into structured scenarios for downstream use.

HM3D: This dataset contains 755 annotated scenes. Since floor heights are not provided, we estimate them by clustering the Y-axis values of camera poses. This enables us to establish clear floor boundaries, which are essential for scene construction and co-visibility graph generation.

B. Annotation generation details

B.1. Co-Visibility annotation overview

As described above, with the 3D assets and depth information provided in the dataset, the ground truth annotation of co-visibility between two images should theoretically be straightforward. However, this process is more complex than it seems. We follow the dataset generation process outlined in Sec. 3.3. First, while placing cameras and selecting images, the image set must remain sparse, maintaining both co-visibility space and maximum scene coverage. Second, as a reasoning task, the annotations should align with human perception while being automatically applicable to large-scale datasets. To ensure annotation accuracy and consistency, we curate a smaller, human-annotated dataset, which serves as a human reasoning baseline for benchmarking Co-VisiON and validating the correctness of the automatically generated dataset.

B.1.1. Automatic co-visibility annotation

Camera pose placing strategy. While we could place cameras anywhere within the navigable areas of a scene, we select positions near walls or furniture to better mimic human photography behavior. This strategy aims to emulate real-world scenarios, where photographers often position themselves near peripheral areas to maximize coverage. As shown in Fig. I, the red-bordered regions between black obstacles and white navigable spaces indicate our preferred camera locations. The cameras are generally oriented away from walls to ensure comprehensive coverage, capturing a wide range of visual features that aid both neural network analysis and human interpretation.

Image selection criterion. First, to assess the scene coverage and the overlapping regions between any two images, we convert the depth image into a point cloud in global coordinates and use it in the subsequent scoring process. We select images in progressive iterations by adopting a scoring method to ensure we cover the entire scene with the fewest possible photos. In each iteration, n_c randomly sampled candidates are generated based on the camera pose placing strategy where every candidate, n_c^i, possesses observations (RGB and Depth) along with the pose information of the agent and its sensors. We then choose the highest-scoring candidate based on the scoring function:

$$S = \alpha \cdot O_u + \beta \cdot O_p, \tag{2}$$

where O_u represents the newly explored region covered by the projected point cloud, while O_p denotes the previously explored region recorded from previous iterations. The α and β are set to 0.9 and 0.1, respectively. This configuration reflects a preference for camera poses that explore more of the uncovered area.

After selecting the best candidate from n_c possible candidates based on the weighted score as in (2), we then remove the positions near the best candidate (or within some radius r) from the viable candidate selection area using equation (3):

$$d(p_i, p_s) <= r \tag{3}$$

where $d(.)$ represents the Euclidean distance between the possible candidate location p_i and the selected best candidate p_s. This particular pruning is shown as light-orange blobs in Fig. I. We observe that it is effective in covering the scene faster and encouraging sparse image sets. We iterate this data generation process until we explore more than 80% of the scene. All observations corresponding to the selected best candidates as well as their pairwise IOU are saved.

B.1.2. Human co-visibility annotation

Besides developing the automatic annotation method, we also have human annotators manually label a subset of

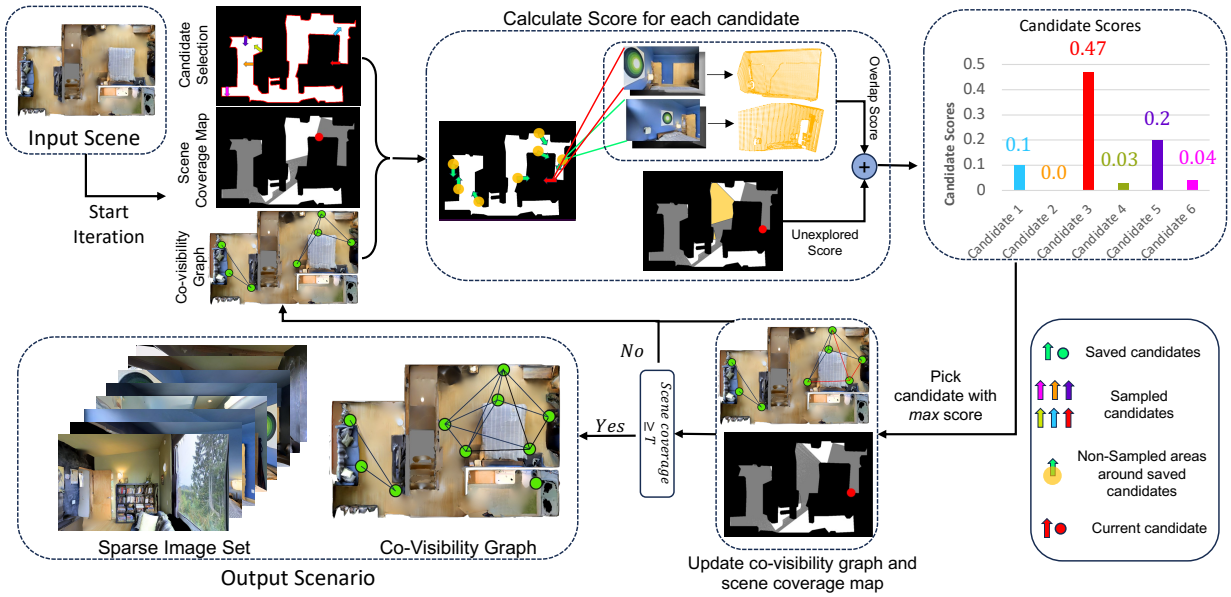


Figure I. **Dataset generation:** Firstly, candidates are sampled from the possible locations shown as red border of the top-down map in candidate selection. Then the sampled candidates, along with a scene coverage map and the current co-visibility graph, are used to calculate scores for each candidate. The scores are computed with the best saved candidates from previous iterations. The top-scoring candidate among the currently sampled is chosen to update scene coverage and the co-visibility graph. This process is repeated until a scene coverage threshold is reached, at which point the best candidates' observations with a co-visibility graph are saved to create the dataset.

Figure II. **We are able to generate different scenarios for a single scene. This is an example of 5 scenarios generated for Scene Goff**

scenes. The human annotation mirrors how humans reason about spatial relationships and, therefore, serves two purposes: it helps assess the quality of automatic annotation, and more importantly, provides a human reasoning baseline for benchmarking Co-VisiON.

The process involves 6 scenes arbitrarily chosen from the automatically generated dataset from Sec. B.1.1. To facilitate human annotation, we develop a website with a GUI that loads pairs of images within the same scene for the trained annotators to determine if they can reason the co-visibility of any pair of images, and click to label them accordingly. Images are uniformly sampled from the same scenario to form a pair and presented to the annotators.

Given images $I_a, I_b \in \mathcal{I}$, the following criteria were considered for each pair:

- **Shared objects**: If I_a and I_b share the view of the same objects such that an annotator can infer the two images' relative pose, they are labeled as connected.
- **Object Continuity**:
  If I_a shows the left side of a sofa and I_b shows the right side, even if the sofa is not fully visible in either image, the partial views align in a way that the viewer can per-

ceive them as continuous, and spatially reason the images as parts of the same object.

- **Sub-Scene Relationship**: one image may be a more zoomed-in portion of the other. Or an image may have details obstructed by objects in its view.
- **Featureless Surface**: If the overlapping region of I_a and I_b is devoid of distinguishable features—such as a plain wall—the pair is labelled as not connected. The absence of features hampers the ability to establish a clear co-visibility relationship between the images.

Along with the above criterion, this manual annotation process undergoes a stringent cross-validation phase where each scene is annotated at least twice by different annotators. When it is challenging to identify relationships in ambiguous scenarios, the images may be marked for further review. The resulting sets of annotations are then compared and discussed until a complete agreement is reached among the annotators. In this way, we ensure the annotations of spatial relationships are as precise and robust as possible.

We will release both the automatically labeled and the manually labeled datasets to facilitate future research. The human annotation website will also be released so that anyone can extend the Co-VisiON task to a larger scale and for more diverse scenarios.

C. Visualization of multiple scenarios from a single scene

We can simulate different scenarios within the same scene by manipulating the seed, resulting in distinct image combinations. In this supplementary, we present five scenarios generated for Scene Goff through this approach shown in Fig. II.

D. Evaluation metrics

As discussed in Sec. 3.2, we use IoU (Intersection over Union) and AUC (Area Under Curve) as the evaluation criteria which are mathematically defined as following.

$$\text{IOU}(\mathbf{A}, \hat{\mathbf{A}}) = \frac{\sum_i \sum_j \mathbf{A}_{ij} \wedge \hat{\mathbf{A}}_{ij}}{\sum_i \sum_j (\mathbf{A}_{ij} \vee \hat{\mathbf{A}}_{ij}) + \epsilon} \tag{4}$$

where $\mathbf{A}_{ij}$ and $\hat{\mathbf{A}}_{ij}$ stand for the binary elements at the ith row and jth column of the corresponding matrices of the ground truth and predicted co-visibility graphs, $\mathcal{G}$ and $\hat{\mathcal{G}}$. The $\wedge$ and $\vee$ represent element-wise AND and OR operations respectively. A small constant ϵ is added to the denominator to avoid zero division.

Another metric that we use is Area Under Curve (AUC). Formally, suppose $\hat{\mathbf{A}}_\tau$ is the binary predicted matrix given the threshold τ, and τ is sampled from a discrete and ordered set $\{t_1, t_2, \ldots, t_n\}$ from range $[0, 1]$. The AUC is computed as:

$$\text{AUC} = \sum_{i=1}^{n-1} \frac{(\text{IOU}(\mathbf{A}, \hat{\mathbf{A}}_{t_{i+1}}) + \text{IOU}(\mathbf{A}, \hat{\mathbf{A}}_{t_i}))}{2} \cdot (t_{i+1} - t_i) \tag{5}$$

E. Details and graph definition in DUSt3R experiment

To evaluate the performance of co-visibility prediction under different graph structures, we experiment with several types of graphs in the DUSt3R benchmark. Each graph is defined as follows:

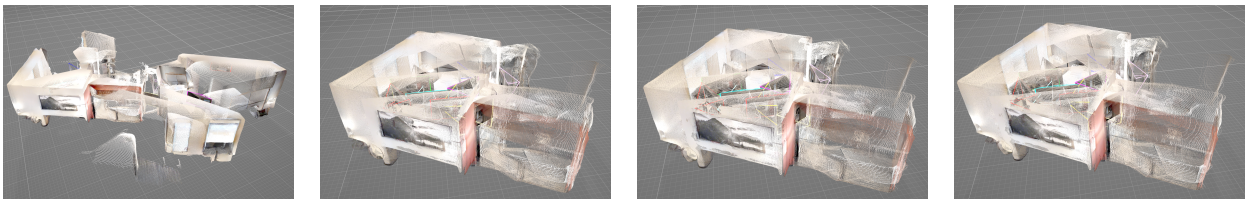
- **Complete graph** refers to a graph where every pair of images is directly connected. However, due to its resource-intensive nature, it becomes less practical for large-scale reconstructions.
- **Co-visibility graph** is resource-efficient and maintains geometric coherence, making it a viable alternative to the Complete strategy.
- **Star graph** consists of a central frame connecting all other images. This approach emphasizes a central perspective and proves valuable when a single viewpoint dominates the scene.
- **Ground Truth graph** connects a pair of images in its graph if they are within a predefined proximity. This graph relies on the absolute pose obtained from the ground truth for each pair of images.

In addition, we present a comparison of these graphs' performance in terms of 3D reconstruction accuracy and computational cost. As shown in Fig. III, we visualize how each graph affects image pairing during reconstruction. The advantages and limitations of each graph type, including aspects such as memory usage and scalability, are discussed in detail in Sec. 5.1. This analysis helps justify our choice of the co-visibility graph as the optimal solution for balancing accuracy and computational efficiency.

F. Detailed definition of different graphs used in CroCo experiment

We train a model to evaluate its performance over the co-visible regions defined by different graph structures, including: (1) High Co-visibility Graph, (2) Co-visibility Graph, and (3) Random Graph.

- **High co-visibility graph**: This graph is created using the pairs of images from a scenario with high spatial and visual overlap ($\geq$50%, or IoU *i*). Using a threshold *i*, it selects pairs with significant overlap, focusing on high-confidence spatial relationships. This graph aims to test the CroCo model's capability to reconstruct a masked target image using a reference image from a different perspective, particularly emphasizing close spatial relationships.



(a) Star - Scene 1 (b) Covision - Scene 1 (c) Complete - Scene 1 (d) GT - Scene 1

(e) Star - Scene 2 (f) Covision - Scene 2 (g) Complete - Scene 2 (h) GT - Scene 2

Figure III. Two examples of 3D reconstruction results by DUSt3R, each based on distinct graphs.

- **Co-visibility graph**: The Co-Visibility graph of a scenario is the same graph as discussed in Sec. 3.1 and Sec. 3.3. This usually contains image pairs with both high and low visual/spatial overlap connected by an edge. From these pairs, we randomly select the same number of edges as in the High Co-visibility Graph. This selection ensures a balanced comparison, containing a mixture of high and low co-visibility pairs. This graph is useful in testing the CroCo model's ability to reconstruct images from a variety of perspectives, leveraging both high and low co-visibility information.
- **Random graph**: The Random graph is constructed by selecting all possible image pairs within a scenario, including those with little to no spatial or visual overlap. A subset of these pairs is randomly selected, ensuring the number of edges matches that of the High Co-visibility Graph. This diverse selection, which includes pairs with minimal or no overlap, tests the CroCo model's ability to handle a variety of image combinations, challenging it to perform well across more unpredictable scenarios.

G. Ablation Study of CoVis

G.1. Zero-shot Evaluation

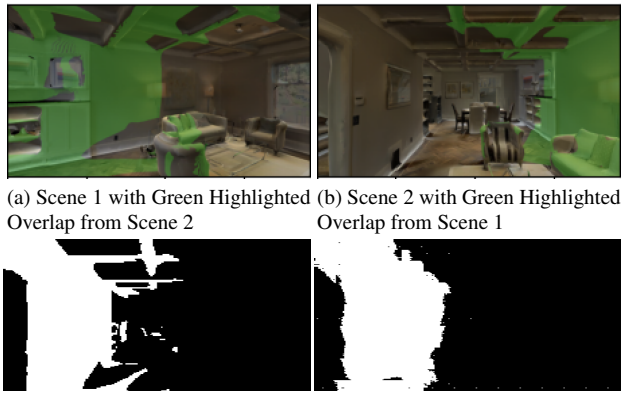
We further evaluate the generalization ability of our Covis by training on one dataset (either Gibson or HM3D) and testing on both Gibson and HM3D individually. As shown in Tab. I, Multi-view Covis consistently outperforms pairwise Covis in both in-domain and zero-shot settings, demonstrating better robustness to dataset's domain shifts. We also evaluate the predicted masks using the standard IoU metric (not graph IoU), and achieve an average IoU of 67.3% against the binarized ground-truth masks.

Table I. Zero-shot evaluation of multi-view Covis and pairwise Covis across different training and test sets.

Train	**Test**	**Multi-view Covis**		**Pairwise Covis**	
		IoU	AUC	IoU	AUC
Gibson	Gibson	0.56	0.52	0.51	0.47
Gibson	HM3D	0.51	0.48	0.48	0.44
HM3D	Gibson	0.58	0.53	0.55	0.50
HM3D	HM3D	0.52	0.48	0.50	0.46

G.2. Qualitative Result of Mask Prediction

We also visualize the ground-truth co-visibility masks between image pairs; which serve as additional supervision signals during training, and the binary co-visibility masks learned by Covis model in Fig. IV.



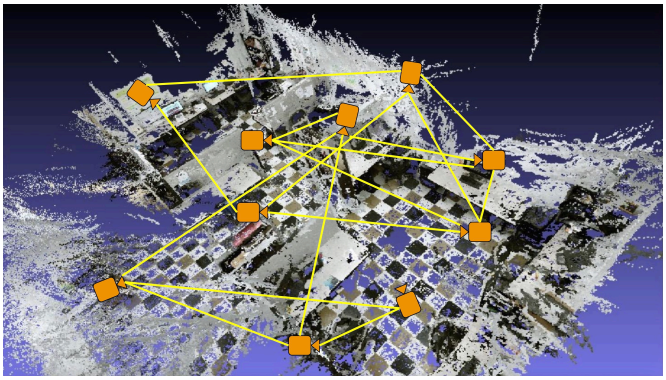
(a) Scene 1 with Green Highlighted Overlap from Scene 2 (b) Scene 2 with Green Highlighted Overlap from Scene 1

(c) Binarized GT mask for (a) (d) Binary predicted mask for (a)

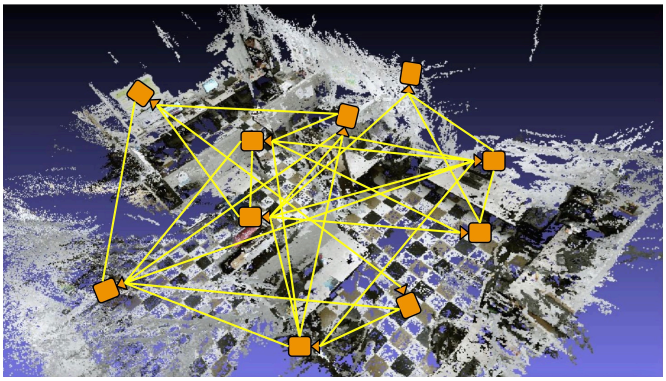
Figure IV. Visualization of the co-visible region between a pair of images (a) and (b), with (c) showing the binarized ground truth mask and (d) the predicted mask on the image from (a).

H. Sim2Real

In this experiment, our aim is to demonstrate that the Co-VisiON dataset exhibits a small domain gap compared to real-world environments, such as the AVD dataset [4]. We have tested Covis, ViT, VGG, Resnet, Contrastive, and NetVlad methods on the AVD dataset using pretrained model from Co-VisiON dataset. We observe AUC metric results in Tab. II to be comparable to those in Tab. 3. Similarly, from Fig. V, we can see that the co-visibility graph predicted using the Covis closely resembles the manually labelled topology graph.



(a) Manually Labeled Topology



(b) Covis Predicted Topology

Figure V. Comparison of manually labeled and Covis predicted Topologies. For clarity, we show 10 sampled images to provide an illustrative example.

Table II. Zero-shot comparison of AUC values (%) for baseline models pretrained on the Co-VisiON dataset and evaluated on real-world data.

Baseline	**Covis**	**ViT**	**VGG**	**ResNet**	**Contrastive**	**NetVlad**
AUC	0.61	0.52	0.33	0.32	0.31	0.22